Language-Universal Speech Processing: Lessons from ASAT and Large Pre-train Models with Extensions to Multilingual ASR

Chin-Hui Lee
School of ECE, Georgia Tech
Atlanta, GA, USA
chl@ece.gatech.edu

Thanks to Prof. S. M. Siniscalchi and Prof .Y. Tsao for ASAT collaboration

1

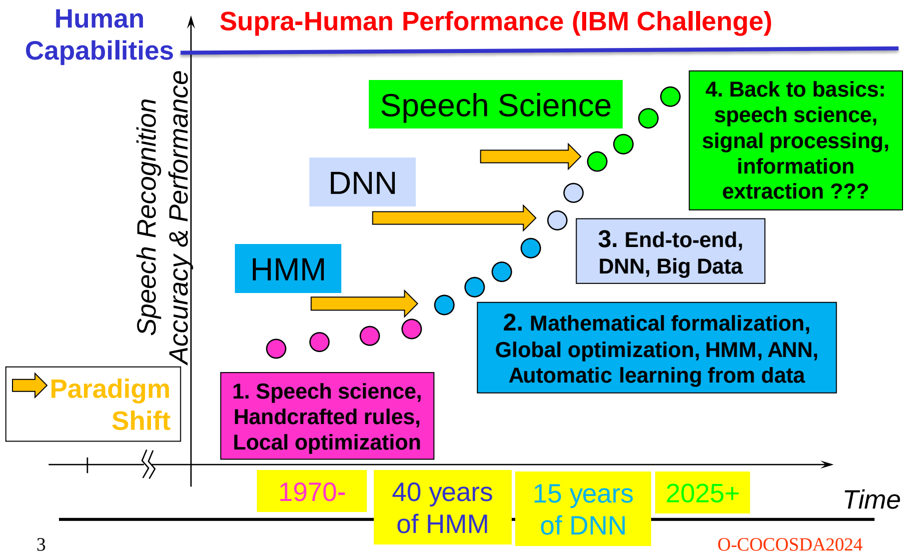
Outline and Talk Agenda

- Brief history of ASR and key messages
- State-of-the-art top-down ASR: blackbox, data-driven
  - Current capabilities and limitations: nice but not good enough
- What more can be done? What's next?
  - Automatic Speech Attribute Transcription (ASAT)
  - Bottom-up attribute detection and knowledge Integration
- Recent ASAT effort: language-universal speech units
  - Attribute-based visible speech and multilingual ASR, etc.
- Recent ASR and DNN efforts
  - Large pre-train models and tools for multilingual ASR
  - Hopfield and Hinton just won 2024 Nobel Prize (Physics)
- Conclusion and future work – everyone can contribute

2 O-COCOSDA2024

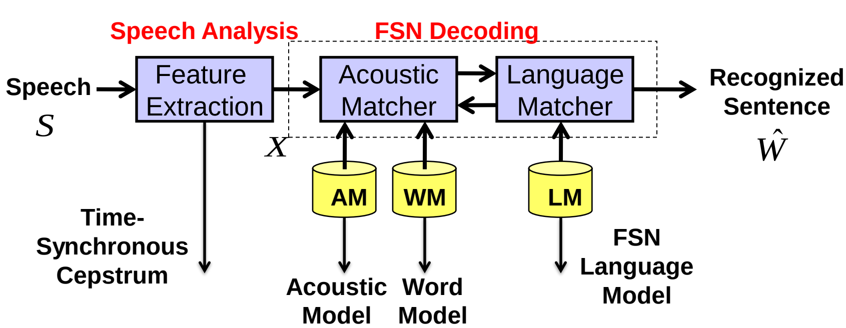
2

A Brief History of Speech Recognition



State-of-the-art ASR Capabilities (1/3)

- Use statistical pattern recognition approaches
  - Rigorous mathematical formulation: e.g., HMM, DP and ANN
  - Acoustic models with tens of million of parameters
  - Language models with hundred of million of parameters
  - Same for all spoken languages: little linguistic knowledge used



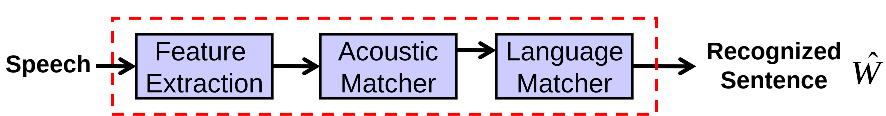
State-of-the-art ASR Capabilities (2/3)

- Use HMM to model phones, words and sentences
- Work well if a task follows some specified training protocols
  - Speaker: speaking rate, accent, age, gender, emotion state, etc.
  - Speaking environment: channel, background noise, etc.
  - Acoustics and signal acquisition devices, push-to-talk, etc.
  - Domain knowledge: vocabulary, syntax, semantics, etc.
- Achieve high accuracies for resource-rich languages
  - English, Mandarin, Arabic, and many others
- Extend ASR learning methodology to other communities, e.g., machine translation, text understanding, bioinformatics
- Deploy many data-driven modeling tools for HMM, ANN, LM
  - But do they lower entry barriers to ASR and advance technologies?

State-of-the-art ASR Capabilities (3/3)

- From HMM to deep neural network (DNN, Hinton)
  - Combining frontend feature extraction and backend scoring
  - Maintaining finite-state network (FSN) search
  - Leveraging on huge amounts of training speech and text, and resulting in large pre-trained models, e.g., Whisper, Nemo, etc.
  - Offering flexible DNN architectures and billions of parameters

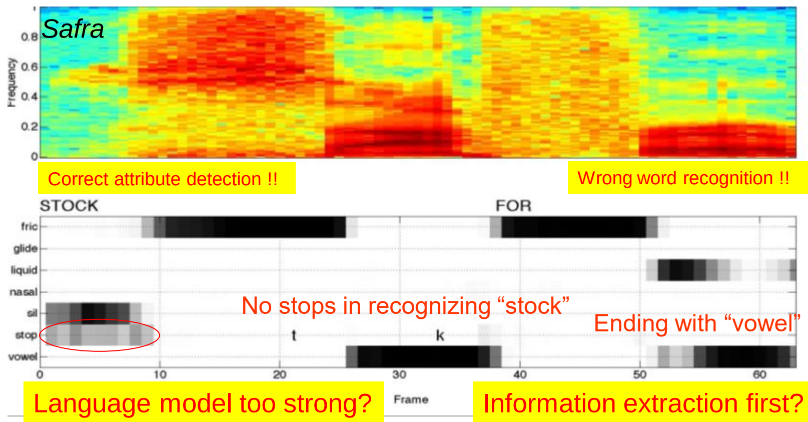
- Use end-to-end modeling (DNN) and FSN decoding



- Work well for resource-rich languages achieving low WERs
  - Fine-tuning to obtain models for resource-limited languages

A Problem with Top-Down Integrated Search

- Manner probability evolution is shown in the location around the error: *Safra* → *Stock For*



From Blackbox Learning to Explainable AI

- Brute-force score-based image and speech recognition (粗功)
- No detailed analysis (細功**)**: requiring domain knowledge for problem solving, not just blind tag-based DNN learning
- What about today's top-down ASR?
  - Giving unexpected results: not human-like natural user interfaces (NUIs)
- From black-box to white-box: knowledge-driven modeling for ASR
  - Automatic speech attribute transcription (Lee, *et al, Proc. IEEE*, 2013)
- Desperately needed new effort: knowledge-driven Explainable AI

Human-Based Speech Processing

- Human speech recognition (HSR): no 'strange' errors
- Learning from spectrogram reading and HSR
  - ➢ Explore speech knowledge hierarchy, from acoustics to pragmatics
  - ➢ Incorporate acoustic and auditory cues in speech
  - ➢ Weigh & combine evidences to form cognitive hypotheses
  - ➢ Verify them until consistent decisions are reached

➡ Bottom-up knowledge source integration
- ➢ But also leveraging upon 55 years of data-driven modeling
- ➢ ASAT: providing a collaborative vehicle

Learning from Speech Science

Vast speech literature & ideas yet to be explored in ASR

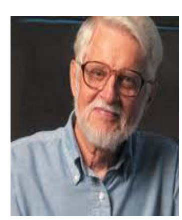
. . .



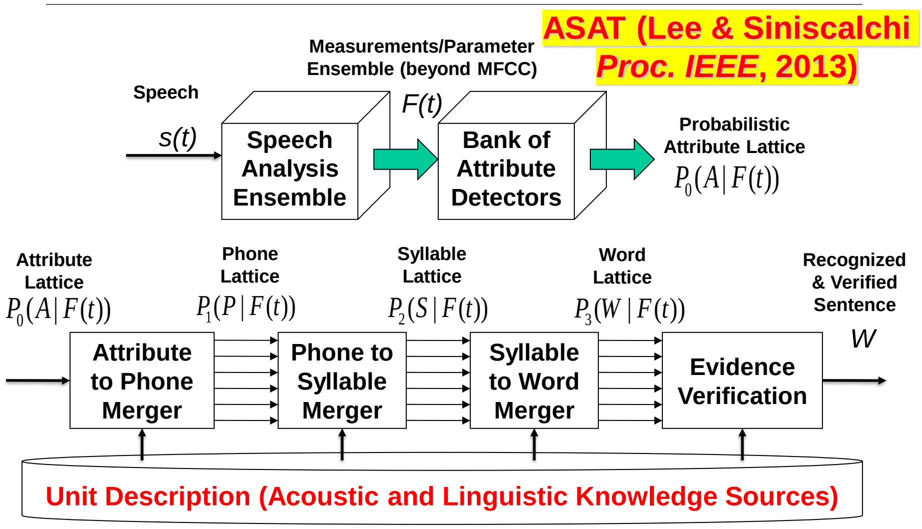

. .

. .

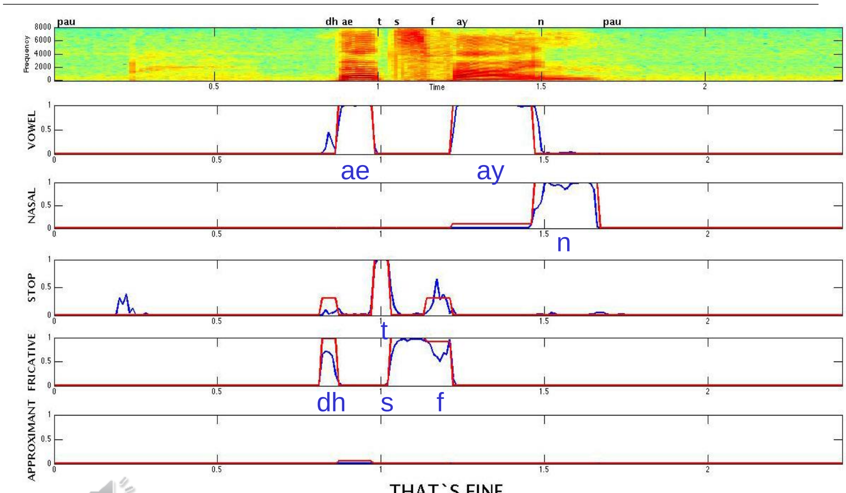



. . .

Automatic Speech Attribute Transcription

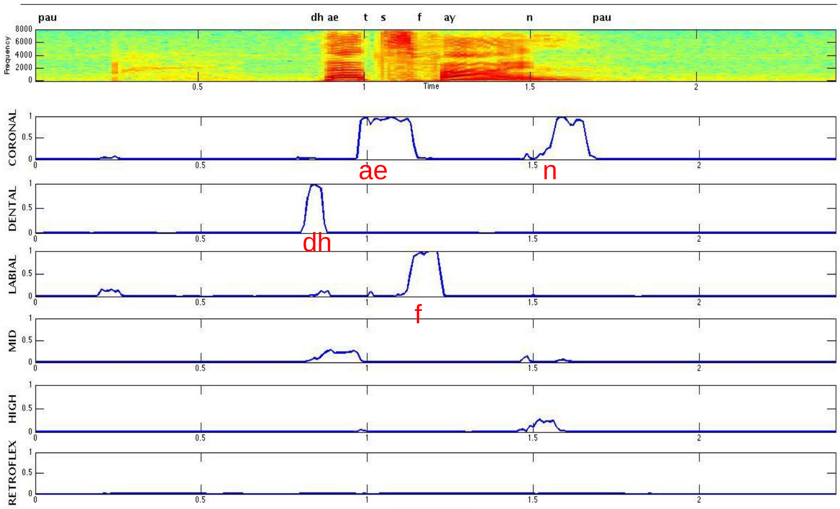


Detection of Manner of Articulation



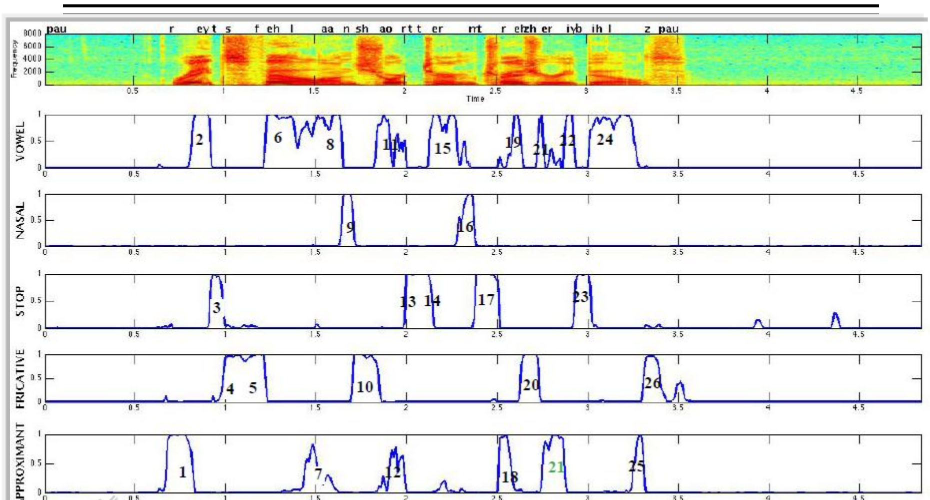
(Plots generated by 3-NN in Blue and HMM in red)

Detection of Place of Articulation



THAT`S FINE

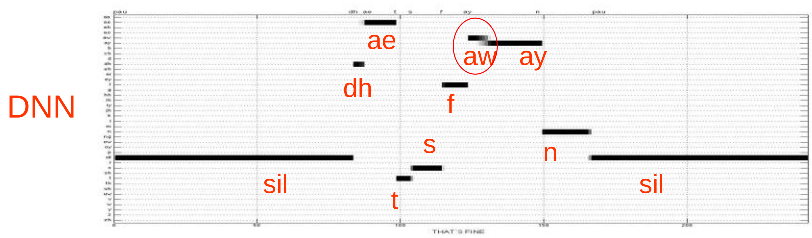
Another Visible Speech: Landmarks



RATES FELL ON SHORT TERM TREASURY BILLS (blue curves generated by DNN)

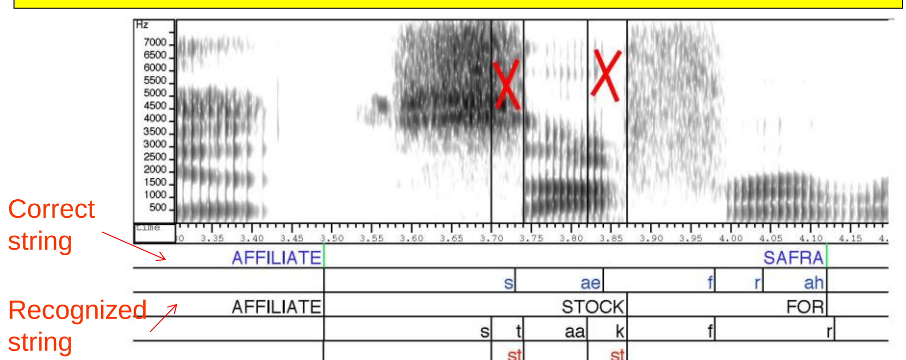
DNN-based Phone Posteriogram

- From HMM to DNN models: better accuracies
- Posteriorgram: DNN outputs simulating posteriors
  - ➢ Clear lines indicating high detection probabilities
  - ➢ “that’s fine” or “sil dh ae s f ay n sil”



Lattice Rescoring for Error Correction

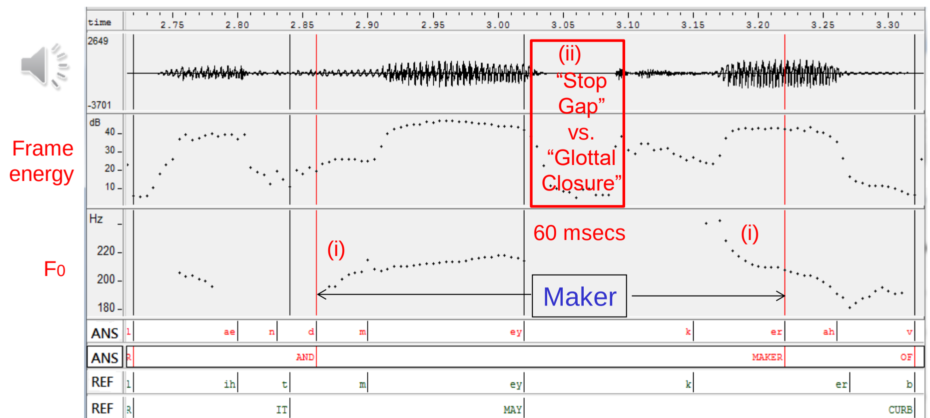
Reduce 30% of word errors in the 30 “worst” utterances



- Penalizing HMM scores with absence of the stop attributes
- Other attribute detectors can function similarly when needed

Suprasegmental Prosody and Duration Features for Correction (Future Paper)

If the Fed pushes the dollar higher, ~~AND MAKER OF~~ IT MAY CURB the demand for U.S. exports.

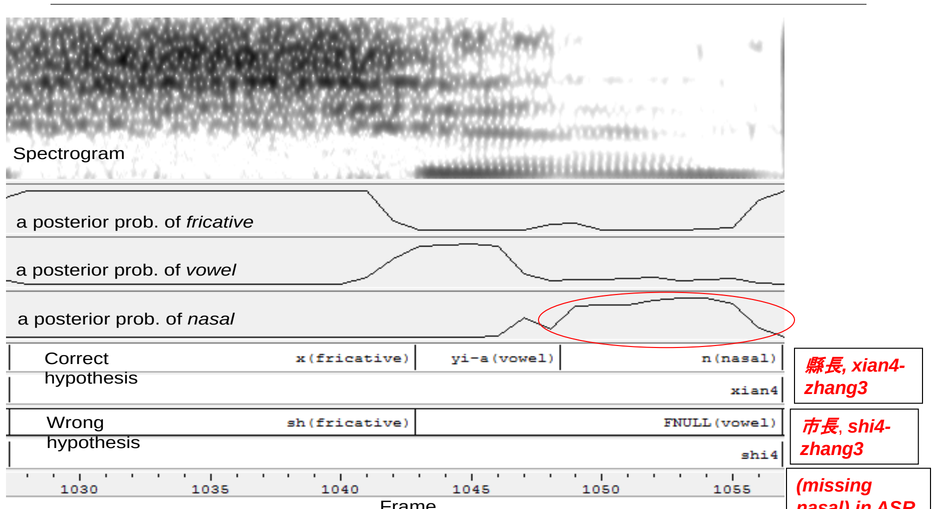


Two inconsistencies: (i) the F0 for the segment "ker" is too high with respect to that for the preceding segment "ma", and (ii) the "glottal closure" of stop sound in "maker" is too long

17 Dr. Chen-Yu Chiang, National Taipei Univ., produced the figure O-COCOSDA2024

17

Language-Universality: American Manner Detectors for Error Correction in Mandarin



18 Thanks to Dr. Chen-Yu Chiang of Taipei University for the figure O-COCOSDA2024

18

Knowledge Integration in Mandarin LVCSR

	WER (%)	CER (%)	SER (%)
Baseline	13.75	10.56	7.79
+Manner	13.45	10.20	7.44
+Break	12.57	9.81	7.13
+M+B	12.43	9.36	6.90
+B+Pitch	12.26	8.93	6.73
+M+B+P	12.24	8.85	6.63

Manner, break and pitch models all improve Mandarin ASR performances progressively & additively (ISCSLP2012)

Attribute Detection Performance

- 21 detectors (ANN, not HMM)
  – WSJ0 training (excluding 6% CV)
  – Nov92 testing (330 utterances)
  – Little context, no lexicon, no syntax

  – Nasal: 97.1%
  – Dental: 99.1%
  – Glottal: 99.7%

  – Tense: 90.50%
  – Continuant: 89.93%
- DNN: all detection rates > 90% (ICASSP2012)

Attribute	ANN	Naïve
anterior	93.2	63.8
back	92.9	80.4
continuant	89.93	55.7
coronal	93.1	74.5
dental	99.1	98.9
fricative	95.4	84.7
glottal	99.7	99.2
approximant	95.9	90.8
high	94.9	83.3
labial	92.5	89.0
low	96.9	90.7
mid	93.6	88.2
nasal	97.1	91.3
retroflex	98.4	93.8
round	93.4	85.3
stop	94.9	84.7
tense	90.5	60.5
velar	98.4	94.6
voiced	95.4	59.9
vowel	91.3	67.5

Multilingual ASR: Current Status

- Recent advances: single large pre-trained models
  - Whisper, Nemo, etc. plus language-specific fine-tuning
- Technology dimensions
  - Modeling unit: language-universal vs language–specific, such as IPA, characters (CTC), speech attributes (ASAT)
  - Word modeling based on acoustic modeling units
  - Language modeling (LM): language-dependent
  - Training data: resource-rich vs resource-limited settings
  - Feature: language-universal vs language–specific
- From domain-specific to domain-independent LM
  - WSJ0 WER: 4% (trigram), 7% (bigram), 70% (0-gram)

Effect of Features on Language Identification

- Our recent study on multilingual ASR (Interspeech2024)
- Multilingual Spoken Words Corpus (MSWC for isolated commands)
  - 8 in-domain (ID) languages with 500 in-vocabulary (IV) training samples
  - 3 in-domain out-of-vocabulary (ID-OOV) languages
  - 3 lout-of domain (OD) out-of-vocabulary unseen language (UL)
  - 30 samples in each evaluation set in each language
- Language ID models trained by language-specific data
- Domain adversarial training (DAT) for language-universal as compared to conventional feature extraction (FE): DAT hurts phone models the most
  - DAT reduces language specificity and greatly degrades ID accuracies

Table 3: *Language identification accuracy (%) with (w/) and without (w/o) DAT for characters, phonemes and attributes.*

System (↓)/Units (→)	Characters	Phonemes	Attributes (ours)
w/o DAT	91.35	91.10	90.47
w/ DAT	45.24	49.66	**34.10**

Spoken Keyword Recognition (SKR)

- In-domain (seen languages), in-vocabulary (ID-IV) SKR
  - Base$_{char}$ and Base$_{Phone}$ use language-specific training
  - Base$_{attr}$ is language-universal and did not perform as well
- After DAT, DAT$_{attr}$ outperforms DAT$_{char}$ and DAT$_{phone}$

Table 4: *Testing WER (%) of the in-domain set on 8 rich-resource languages and the average (Avg.).*

System	ID-IV								
	en	de	fr	fa	es	ru	it	pl	Avg.
Base$_{char}$	13.14	12.57	13.65	14.77	11.89	15.50	**14.39**	15.00	13.86
Base$_{phone}$	**12.73**	**11.70**	**13.29**	14.30	**11.84**	**12.96**	14.63	15.14	**13.32**
Base$_{attr}$ (ours)	13.28	12.04	13.69	**13.53**	12.20	13.67	15.13	**14.98**	13.56
DAT$_{char}$	18.73	16.45	19.55	16.91	14.99	17.98	17.19	18.62	17.55
DAT$_{phone}$	17.76	18.33	21.12	20.00	17.58	14.97	18.91	17.65	18.29
DAT$_{attr}$ (ours)	**15.51**	**13.74**	**16.27**	**15.47**	**13.85**	**13.83**	**16.73**	**15.75**	**15.14**

In-Domain Out-of-Vocabulary (ID-OOV) SKR

- Models obtained as before (no retraining, zero-shot transfer)
  - Base$_{char}$ degrades greatly (poor character sequence modeling)
  - Base$_{attr}$ & Base$_{Phone}$ are language-consistent and perform better
- After DAT, slight improvements are observed

Table 5: *Zero-shot transfer: Testing WER (%) of the 3 in-domain out-of-vocabulary (**ID OOV**) keywords from Russian, Italian, and Polish.*

System	ID OOV			
	ru	it	pl	Avg.
Base$_{char}$	63.96	44.50	40.77	49.74
Base$_{phone}$	**31.15**	**40.62**	33.29	**35.02**
Base$_{attr}$ (ours)	31.57	41.89	**32.80**	35.42
DAT$_{char}$	54.89	40.19	37.95	44.34
DAT$_{phone}$	**28.15**	**38.61**	34.11	**33.62**
DAT$_{attr}$ (ours)	29.98	40.28	**30.23**	33.81

SKR of Unseen Languages (Phone Mismatch)

- OOV in 3 unseen languages: Turkish, Latvian, Lithuanian
- Models obtained as before (no retraining, zero-shot transfer)
  - $Base_{attr}$ outperforms $Base_{char}$ and $Base_{Phone}$ ($Base_{char}$ is the worst)
- After DAT, $Base_{attr}$ performs even better (the best so far)

Table 6: *Zero-shot transfer: Testing WER (%) of the 3 unseen languages (**UL**), namely Turkish, Latvian, and Lithuanian.*

System	UL			
	tr	lv	lt	Avg.
$Base_{char}$	61.79	61.18	39.39	54.12
$Base_{phone}$	54.50	45.50	43.64	47.88
$Base_{attr}$ (ours)	**48.33**	**40.10**	**30.30**	**39.58**
DAT_{char}	61.23	57.33	45.45	54.67
DAT_{phone}	46.67	39.59	49.70	46.33
DAT_{attr} (ours)	**42.96**	**36.25**	**26.36**	**37.10**

25

25

SKR with Large Pre-trained Models

- End-to-end (E2E) models: Google, Facebook, Meta, Nvidia, etc.

➡ Replying on LM for resource-rich languages in multilingual ASR

- Average SKR WER comparisons for 8 seen languages (no LM)

ID-IV ($Base_{phone}$)	Whisper (Fine-tuned)	Facebook (Fine-tuned)
13.32%	56.73% (13.54%)	54.39% (xx%)

➡ Large pre-trained models do not do much better after fine-tuning, but require much bigger model sizes (95 MB vs. small Whisper of 244 MB)

===

- Average WERs of OOV for 3 of the 8 seen and 3 unseen languages

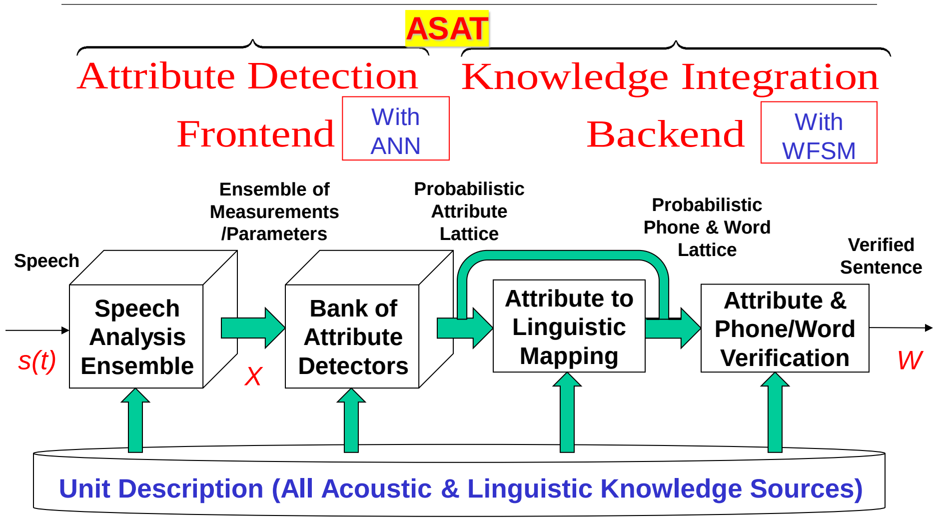
ID-OOV	Whisper (Facebook)	OD-OOV	Whisper (Facebook)
35.02%	77.78% (72.18%)	47.88%	78.64% (73.21%)

➡ With no fine-tuning, large pre-trained models perform much worse

26

26

Next: Multilingual LVCSR and Bottom-Up Keyword Spotting in Extraneous Speech



Conclusion and Future Work

- *Knowledge-ignorant* modeling for pattern recognition is mathematically well-formulated: carrying us a long way so far
- *Knowledge-rich* modeling leverages on data-driven modeling
  - ➢ From top-down decoding to highly-parallel, bottom-up processing
  - ➢ Information integration within the speech knowledge hierarchy
- *Robust information extraction* supplements pattern matching with/plus signal processing to detect "islands of reliability"
  - ➢ A collaborative community effort: everyone can help
- *Final grand challenge: language-universal* modeling
  - ➢ Can we train ASR models for all languages once and for all ??
  - ➢ How do we learn from human language acquisition ??